JURNAL INKUIRI
ISSN: 2252-7893, Vol 3, No. II, 2014 (hal 60-74)
http://jurnal.fkip.uns.ac.id/index.php/sains

PENGEMBANGAN INSTRUMEN EVALUASI *TWO-TIER MULTIPLE CHOICE QUESTION* UNTUK MENGUKUR KETERAMPILAN BERPIKIR TINGKAT TINGGI PADA MATERI KINGDOM *PLANTAE*

Mufida Nofiana[1], Sajidan[2] dan Puguh[3]

[1] Program Studi Pendidikan Sains Program Pascasarjana
Universitas Sebelas Maret Surakarta, 57126, Indonesia
mufida.nofiana@yahoo.co.id

[2] Program Studi Pendidikan Sains Program Pascasarjana
Universitas Sebelas Maret Surakarta, 57126, Indonesia
adjids2002@yahoo.com

3 Program Studi Pendidikan Sains Program Pascasarjana
Universitas Sebelas Maret Surakarta, 57126, Indonesia
karyarina@yahoo.com

Evaluasi merupakan alat yang digunakan untuk mengukur tujuan pembelajaran yang salah satunya adalah penguasaan keterampilan berpikir tingkat tinggi. Penguasaan keterampilan berpikir tingkat tinggi pada materi *kingdom plantae* membutuhkan kemampuan seperti menganalisis, mengevaluasi, dan menciptakan. Pengukuran kemampuan berpikir tingkat tinggi pada materi *kingdom plantae* dapat dilakukan dengan instrumen evaluasi *two-tier multiple choice question*. Penelitian pengembangan instrumen evaluasi bertujuan untuk mengetahui (1) karakteristik instrumen evaluasi *two-tier multiple choice question* yang mampu mengukur keterampilan berpikir tingkat tinggi, (2) kelayakan instrumen evaluasi *two-tier multiple choice question* sebagai evaluasi formatif, (3) respon siswa terhadap instrumen evaluasi *two-tier multiple choice question* yang diterapkan di SMA. Penelitian pengembangan instrumen evaluasi menggunakan model *Research and Development* (R&D) mengacu pada Borg and Gall (1983) yang telah dimodifikasi. Sampel pengembangan meliputi 4 validator ahli, 20 siswa pada uji terbatas, 64 siswa pada uji lapangan, dan 64 siswa pada uji korelasi penggunaan instrumen evaluasi. Analisis data dilakukan dengan rumus persentase dan uji korelasi pearson. Hasil penelitian menunjukkan (1) karakteristik instrumen evaluasi *two-tier multiple choice question* antara lain dikembangkan berdasarkan indikator keterampilan berpikir tingkat tinggi Anderson dan Krathwohl (2001) meliputi menganalisis, mengevaluasi, dan menciptakan; memiliki validitas dengan interpretasi minimal "cukup"; dan reabilitas "tinggi" (2) kelayakan produk instrumen evaluasi dijamin melalui validitas isi yang "baik"; validitas konstruk yang "baik"; validitas butir soal dengan interpretasi minimal "cukup", tingkat kesukaran soal dengan proporsi 15% mudah: 80% sedang: 5% sulit; daya pembeda soal dengan interpretasi minimal "cukup", dan kepraktisan penggunaan yang "baik" (3) respon siswa terhadap penerapan instrumen evaluasi didapatkan melalui uji korelasi antara instrumen bentuk *two-tier multiple choice question* dengan bentuk *multiple choice question*. Terdapat korelasi antara kedua bentuk instrumen tersebut dengan nilai sebesar 0,15. Artinya siswa memberikan respon yang sama dalam mengerjakan kedua bentuk soal tersebut. Instrumen evaluasi *two-tier multiple choice question* dapat digunakan sebagai alternatif instrumen evaluasi formatif di skeolah dengan penerapan lebih lanjut.

Kata kunci: ***two-tier multiple choice question*, keterampilan berpikir tingkat tinggi, *kingdom plantae*.**

PENDAHULUAN

Penelitian Pemetaan dan Pengembangan Mutu Pendidikan (PPMP) di beberapa kabupaten atau kota yang tersebar di propinsi Jawa Tengah berhasil memetakan Standar Kompetensi dan Kompetensi Dasar yang tidak dikuasai peserta didik SMA pada UN tahun 2009 dan 2010 (Sajidan, 2012). Hasil UN Tahun 2009 dan 2010 untuk mata pelajaran Biologi menunjukkan masih terdapat siswa yang tidak tuntas pada materi lumut dan paku (Data UN Puspendik, 2010). Materi lumut dan paku merupakan bagian dari Kompetensi Dasar mendeskripsikan ciri-ciri divisio dalam dunia tumbuhan dan peranannya bagi kelangsungan hidup di bumi (BSNP, 2006). Soal-soal dalam UN adalah soal dengan tingkat kesulitan yang lebih tinggi (*higher order thingking*) dibandingkan dengan soal yang biasa digunakan guru di sekolah.

Laporan PPMP menyebutkan ketidaktuntasan siswa pada Kompetensi Dasar UN salah satunya disebabkan karena soal-soal yang digunakan guru di sekolah masih sangat standar dan tidak memberdayakan keterampilan berpikir tingkat tinggi siswa (Sajidan, 2012). Analisis ketuntasan Kompetensi Dasar pada UN Tahun 2009 dan 2010 dilanjutkan dengan analisis kebutuhan sekolah yang dilaksanakan berdasarkan delapan Standar Nasional Pendidikan (SNP) meliputi standar isi, standar proses, standar kompetensi lulusan, standar pendidik dan tenaga kependidikan, serta standar penilaian (Badan Akreditasi Nasional, 2012). Hasil analisis delapan SNP menunjukkan masih terdapat kelemahan pada pemenuhan standar penilaian di sekolah, terbukti dengan instrumen evaluasi formatif yang digunakan guru di sekolah hanya Taksonomi Bloom tingkat rendah. Persentase penggunaan ranah kognitif Taksonomi Bloom dalam soal yang digunakan guru adalah 30% hafalan (C1), 60% pemahaman (C2), dan 10% analisis (C4), sedangkan soal yang menuntut aplikasi (C3), evaluasi (C5), dan menciptakan (C6) tidak diberikan oleh guru (Bank Soal Biologi SMA 3 Surakarta, 2012). Idealnya tes formatif yang dilaksanakan oleh guru 80% harus mencakup keterampilan berpikir tingkat tinggi (C4-C6) (Standar Penilaian BAN, 2012).

Tes formatif yang sering digunakan guru di sekolah adalah *traditional assessment* (tes tulis) dalam bentuk pilihan ganda (*multiple choice*). Soal pilihan ganda digunakan karena penilaian soal lebih objektif dan penskorannya mudah, tetapi kemungkinan siswa untuk menebak jawaban atau menjawab soal secara untung-untungan sangat besar. Soal pilihan ganda juga kurang mampu mengukur kemampuan kognitif yang lebih tinggi (Purwanto, 2010). Hasil analisis kebutuhan di sekolah mendapatkan kesimpulan bahwa guru membutuhkan instrumen evaluasi yang mampu mengukur keterampilan berpikir tingkat tinggi.

Pentingnya penguasaan keterampilan berpikir tingkat tinggi terdapat dalam beberapa poin Standar Kompetensi Lulusan Sekolah Menengah. Poin yang diharapkan yaitu siswa dapat membangun dan menerapkan informasi atau pengetahuan secara logis, kritis, kreatif, dan inovatif; menunjukkan kemampuan berpikir logis, kritis, kreatif, dan inovatif dalam pengambilan keputusan;serta menunjukkan kemampuan menganalisis dan memecahkan masalah kompleks (Permendiknas No 23 Tahun 2006). Pembelajaran Biologi merupakan pembelajaran sains yang memerlukan kegiatan penyelidikan atau eksperimen sebagai bagian dari kerja ilmiah. Kerja ilmiah menekankan peserta didik untuk berpikir kreatif, kritis, analitis, dan divergen (BSNP, 2006). Kemampuan

peserta didik untuk berpikir kritis dan kreatif termasuk dalam bentuk keterampilan berpikir tingkat tinggi.

Keterampilan berpikir tingkat tinggi merupakan suatu keterampilan berpikir yang tidak hanya membutuhkan kemampuan mengingat, tetapi membutuhkan kemampuan lain yang lebih tinggi. Lewis dan Smith (1993) mendefinisikan keterampilan berpikir tingkat tinggi (*The Higher Order Thinking Skills*) sebagai keterampilan berpikir yang terjadi ketika seseorang mengambil informasi baru dan informasi yang sudah tersimpan dalam ingatannya, selanjutnya menghubungkan informasi tersebut dan menyampaikannya untuk mencapai tujuan atau jawaban yang dibutuhkan. King, *et al* (2010) mengatakan keterampilan berpikir tingkat tinggi pada siswa dapat diberdayakan dengan memberikan masalah yang tidak biasa dan tidak menentu seperti pertanyaan atau dilema, sehingga penerapan yang sukses dari kemampuan ini adalah ketika siswa berhasil menjelaskan, memutuskan, menunjukkan, dan menghasilkan penyelesaian masalah dalam konteks pengetahuan dan pengalaman.

Konsep berpikir tingkat tinggi diturunkan dari Taksonomi Bloom. Sistem ini mengidentifikasi kemajuan yang hierarki dalam menggolongkan tingkatan proses berpikir menjadi tinggi dan rendah. Ada enam tingkatan taksonomi Bloom yakni: pengetahuan, pemahaman, aplikasi, sintesis, dan evaluasi. Tingkatan pertama dan kedua dari taksonomi Bloom dianggap sebagai kemampuan berpikir tingkat rendah, sedangkan empat tingkatan lainnya digolongkan sebagai keterampilan berpikir tingkat tinggi (Miller, 1990 dalam Ball dan Garton, 2005).

Anderson dan Krathwohl (2001) telah merevisi penggunaan Taksonomi Bloom sebagai kerangka konseptual untuk penelitian keterampilan berpikir tingkat tinggi. Pohl (2000) mengungkapkan bahwa dalam Taksonomi Bloom revisi keterampilan yang melibatkan analisis, evaluasi, dan mencipta dianggap sebagai keterampilan berpikir tingkat tinggi. Anderson dan Krathwohl (2001) menyatakan bahwa indikator untuk mengukur keterampilan berpikir tingkat tinggi meliputi menganalisis, mengevaluasi, dan menciptakan.

Indikator untuk mengukur keterampilan berpikir tingkat tinggi meliputi kemampuan menganalisa, mengevaluasi, dan menciptakan (Anderson dan Krathwohl, 2001). *Output* siswa yang memiliki keterampilan berpikir tingkat tinggi tidak hanya dikembangkan dalam proses pembelajaran, tetapi juga harus didukung dengan evaluasi atau tes yang mencerminkan keterampilan berpikir tingkat tinggi karena evaluasi atau tes merupakan bagian yang menyatu dengan pembelajaran di kelas. Evaluasi dapat digunakan untuk mengukur keberhasilan pencapaian indikator pembelajaran yang dilakukan (Arikunto, 2007). Indikator pembelajaran dapat berupa indikator kognitif produk, kognitif proses, psikomotorik, dan afektif. Evaluasi yang digunakan untuk mengukur keterampilan berpikir tingkat tinggi didasarkan pada indikator kognitif produk.

Instrumen evaluasi yang mengukur keterampilan berpikir tingkat tinggi dapat menggunakan berbagai tipe penilaian seperti *modified multiple choice*, konstruksi jawaban singkat, dan konstruksi jawaban panjang seperti yang telah dilakukan oleh Ramirez dan Ganaden (2008). Salah satu alternatif *Modified multiple choice* yang dapat digunakan untuk mengukur keterampilan berpikir tingkat tinggi adalah bentuk *two-tier multiple choice question* (pilihan ganda bertingkat). Bentuk soal *two-tier multiple choice question* dikembangkan oleh Treagust (2006). Treagust menggunakan soal pilihan ganda

bertingkat untuk mendiagnosis kemampuan siswa memahami konsep IPA. Bentuk soal terdiri dari dua tingkatan soal, tingkatan pertama merupakan isi soal yang memiliki dua alternatif jawaban dan tingkatan kedua merupakan alasan jawaban yang dipilih atas dasar pilihan pertama. Pengembangan instrumen evaluasi *two-tier multiple choice question* dilakukan dengan mengaitkannya pada materi *kingdom plantae*.

Materi *Kingdom plantae* merupakan materi yang dekat dengan siswa. Contoh nyata dari materi *kingdom plantae* sering dijumpai di lingkungan sekitar, seharusnya siswa dapat menguasai materi tersebut dengan baik namun pada kenyataanya masih terdapat siswa yang tidak tuntas terutama pada soal-soal *kingdom plantae* yang menuntut keterampilan berpikir tingkat tinggi (Data UN Tahun 2009 dan 2010). Penilaian keterampilan berpikir tingkat tinggi menggunakan *two-tier multiple choice question* pada materi *kingdom plantae* diharapkan mampu melatihkan siswa untuk memberdayakan keterampilan berpikir tingkat tinggi pada materi tersebut.

Halaydina dan Downing (1989) serta Treagust (2006) mengemukakan keunggulan bentuk soal *two-tier multiple choice question,* salah satunya digunakan untuk tujuan tes yang mengukur kemampuan kognitif siswa pada level yang lebih tinggi (*Higher Order Thinking).* Bentuk soal *two-tier multiple choice question* dapat digunakan untuk membantu menguji pemahaman siswa serta membantu mengidentifikasi miskonsepsi yang mungkin dimiliki oleh siswa. Cullinane (2011) menggemukakan penyertaan alasan pada tingkatan kedua dari bentuk soal *two-tier multiple choice question* dapat digunakan untuk meningkatkan keterampilan berpikir tingkat tinggi dan melihat kemampuan siswa dalam memberi alasan. Penyertaan alasan pada tingkatan kedua soal ini dapat digunakan untuk mengurangi terjadinya untung-untungan yang sering menjadi kelemahan dari bentuk soal pilihan ganda biasa. Penilaian soal yang objektif, mudah, dan cepat menjadi keunggulan *two-tier multiple choice question* dibandingkan dengan soal keterampilan berpikir tingkat tinggi yang lainnya contohnya soal essay. Kelemahan dari soal *two-tier multiple choice question* yaitu tidak mampu digunakan untuk mengukur kemampuan verbal siswa seperti soal *essay*.

Metode Penelitian

Penelitian dilaksanakan di SMA Negeri 3 Surakarta dan SMA Negeri 1 Gemolong. Sampel *yang* digunakan adalah siswa kelas X semester genap Tahun Pelajaran 2012/ 2013 berjumlah 149 siswa. Jenis penelitian yang digunakan adalah *Research and Development (R&D).* Prosedur penelitian dimodifikasi dari model pengembangan Borg & Gall (1983) dan dilakukan hanya sampai pada tahap ketujuh.

Tahapan penelitian dan pengembangan meliputi 1) *research and information collecting,* yang dilakukan antara lain mengenali pemasalahan yang ada di lapangan, analisis proses pembelajaran guru, analisis hasil UN, analisis kurikulum, analisis bank soal, dan studi pustaka; 2) *planning*, yang dilakukan antara lain menentukan Kompetensi Dasar materi yang akan dikembangkan, merumuskan tujuan pengembangan dan indikator keterampilan berpikir tingkat tinggi; 3) *develop preliminary from of product,* yang dilakukan antara lain menyiapkan materi pem-belajaran, membuat kisi-kisi soal, mengembangkan produk awal instrumen evaluasi; 4) *preliminary field testing*, yang dilakukan antara lain validasi produk ke ahli dan guru senior, uji skala terbatas kepada 20 orang siswa yang terdiri dari 6

siswa untuk uji satu-satu dan 14 orang siswa untuk uji skala kecil; 5) *main product revision*, yang dilakukan antara lain perbaikan sesuai dengan saran-saran dari hasil *preliminary field testing*; 6) *main field testing*, yang dilakukan antara lain menguji produk pengembangan dalam skala lebih luas pada 64 orang siswa di SMA Negeri 1 Gemolong untuk melihat validitas, reliabilitas, tingkat kesukaran dan daya beda soal; 7) *operational product revision,* yang dilakukan antara lain revisi produk berdasarkan saran-saran dari hasil *main field testing*. Hasil tahap *operational product revision* adalah produk final instrumen evaluasi *two-tier multiple choice question*. Produk final instrumen evaluasi *two-tier multiple choice question* selanjutnya dikorelasikan dengan instrumen evaluasi kontrol bentuk *multiple choice question* pada 64 orang siswa di kelas X.2 dan X.5 SMA Negeri 3 Surakarta.

Instrumen pengambilan data yang digunakan meliputi angket untuk analisis kebutuhan, lembar *check list* 8 SNP, lembar *check list* ketuntasan KD, dokumentasi silabus dan RPP guru, bank soal guru, lembar *check list* penilaian produk, dan lembar *check list* kepraktisan soal. Data analisis kebutuhan dianalisis dengan statistik deskriptif. Hasil angket dideskripsikan untuk menganalisis kebutuhan pengembangan. Hasil dari analisis digunakan untuk mem-pertimbangkan kebutuhan pengembangan instrumen evaluasi . Data penilaian ahli dan guru senior terhadap soal dianalisis dengan teknik deskriptif persentase (Purwanto, 2010). Analisis data dilakukan dengan cara menghitung skor yang dicapai dari seluruh aspek yang dinilai kemudian menghitungnya dengan rumus sebagai berikut:

$$N = \frac{k}{Nk} \times 100\%$$

Keterangan :
N : persentase kelayakan aspek
k : skor hasil pengumpulan data
Nk :skor maksimal (skor kriteria tertinggi x jumlah aspek x jumlah validator)

Tabel 1. Kriteria Interpretasi Skor Validasi Ahli

Interval kriteria	Kriteria	Konversi
86 % ≤N < 100%	Sangat baik	A
72 % ≤N < 85%	Baik	B
58 % ≤N < 71%	Cukup	C
44 % ≤N < 57%	Kurang	D
N ≤ 44 %	Sangat kurang	E

Data penilaian siswa pada uji coba terbatas (uji coba satu-satu dan uji coba kelompok kecil) dianalisis dengan teknik deskriptif persentase (Purwanto, 2010). Perhitungan data yang dilakukan sama dengan perhitungan pada data validasi ahli dan guru senior. Pada uji coba lapangan terdapat dua jenis data, yakni data kualitatif dan data kuantitatif. Data kualitatif diperoleh dari data kepraktisan soal. Data kualitatif diperoleh dari guru pengguna. Data kepraktisan soal dianalisis dengan teknik deskriptif persentase (Purwanto, 2010). Perhitungan data yang dilakukan sama dengan perhitungan pada data validasi ahli dan guru senior, sedangkan data kuantitatif diperoleh dari pengujian soal meliputi uji validitas, reliabilitas, daya beda, dan tingkat kesukaran soal.

Pengujian validitas dilakukan dengan menggunakan Microsoft excel 2007. Validitas instrumen tes tertulis dapat ditentukan dengan menggunakan rumus korelasi. Rumus korelasi yang digunakan adalah rumus korelasional product moment dari Pearson. Pengujian reliabilitas dilakukan Micosoft Excel 2007. Rumus yang digunakan adalah rumus *Alpha-Cronbach*. Tingkat kesukaran soal dihitung melalui proporsi jawaban keseluruhan siswa yang menjawab benar pada soal tersebut. Daya pembeda dihitung melalui selisih jawaban antara proporsi kelompok tinggi yang

menjawab benar dengan proporsi kelompok rendah yang menjawab benar.

Data uji perbandingan instrumen evaluasi *Two-Tier Multiple Choice Question* (TTMCQ) dengan instrumen *Multiple Choice Question (*MCQ*)* dihitung dengan uji korelasi Pearson menggunakan PASW Statistik 18. Pengujian instrumen evaluasi dilakukan pada siswa yang sebelumnya telah mendapat materi kingdom plantae. Hasil pengembangan instrumen diharapkan akan menghasilkan produk yang mampu memperbaiki kualitas soal pilihan ganda dan memperkaya khazanah soal-soal biologi di SMA khususnya pada materi kingdom *plantae.*

Hasil Penelitian dan Pembahasan

Data yang diperoleh dalam penelitian pengembangan antara lain data analisis kebutuhan, data validasi ahli dan praktisi, data hasil uji coba terbatas dan data hasil uji coba lapangan. Data analisis kebutuhan meliputi tingkat pemenuhan standar nasional pendidikan (SNP) di SMA Negeri 1 Gemolong dan SMA Negeri 3 Surakarta, analisis bank soal biologi yang digunakan oleh guru di sekolah, dan wawancara.

Tabel 2. Hasil Pemenuhan SNP di SMA Negeri 3 Surakarta

Tabel 3 Hasil Pemenuhan SNP di SMA Negeri 1 Gemolong

Tabel 2 dan Tabel 3 menunjukkan tingkat pemenuhan 8 SNP di SMA Negeri 1 Gemolong dan SMA Negeri 3 Surakarta. Analisis hasil pemenuhan delapan SNP di SMA Negeri 3 Surakarta dan SMA Negeri 1 Gemolong menunjukkan bahwa tingkat pemenuhan SNP di masing-masing sekolah termasuk dalam kategori sangat baik, namun pada standar proses dan standar penilaian masih terdapat GAP yang cukup besar antara skor di lapangan dengan skor ideal. Standar proses dan standar penilaian pada kedua SMA perlu mendapat perhatian untuk ditingkatkan.

Wawancara dengan guru menunjukkan bahwa dalam proses pembelajarannya guru tidak terbiasa melatih siswa untuk memberdayakan kemampuan berpikir tingkat tinggi. Analisis bank soal yang digunakan guru dilakukan untuk mengetahui persentase penggunaan tingkat taksonomi Bloom dalam soal. Hasil temuan bank soal guru disajikan pada Tabel 4.

Tabel 4. Persentase Penggunaan Taksonomi Bloom pada Soal di Sekolah

Tingkat taksonomi Bloom	Jumlah soal	Total soal	Persentase (%)
C1 (pengetahuan)	30	100	30
C2 (pemahaman)	60	100	60
C3 (aplikasi)	0	100	0
C4 (analisis)	10	100	10
C5 (evaluasi)	0	100	0
C6 (mencipta)	0	100	0

Tabel 4 menunjukkan bahwa sebagian soal guru masih belum memberdayakan kemampuan berpikir tingkat tinggi (C4 - C6). Padahal idealnya 80% soal yang digunakan guru di sekolah mencakup C4-C6 (BAN, 2006).

SNP	Jml indikator	Skor ideal	Kontribusi	Implementasi SNP		GAP
			%	Skor	%	%
I	8	24	11, 11	21	9, 72	1, 39
II	10	30	13, 89	22	10, 19	3, 70
III	12	36	16, 67	30	16, 67	0
IV	11	33	15, 28	30	13, 89	1, 39
V	11	33	15, 28	31	14, 35	0, 93
VI	4	12	5, 56	12	5, 56	0
VII	3	9	4, 17	9	4, 17	0
VIII	13	39	18, 06	32	14, 81	3, 25
Tot	72	216	100	193	89, 35	10,8

SNP	Jml indikator	Skor ideal	Kontribusi	Implementasi SNP		GAP
			%	Skor	%	%
I	8	24	11, 11	21	9, 72	1, 39
II	10	30	13, 89	20	9, 26	4, 63
III	12	36	16, 67	30	13, 89	2, 78
IV	11	33	15, 28	30	13, 89	1, 39
V	11	33	15, 28	30	13, 89	1, 39
VI	4	12	5, 56	12	5, 56	0
VII	3	9	4, 17	9	4, 17	0
VIII	13	39	18, 06	27	12, 50	5, 56
Tot	72	216	100	178	82, 41	17, 6

Keterampilan berpikir tingkat tinggi (*higher order thinking skill*) adalah keterampilan yang terjadi ketika seseorang

mengambil informasi baru dan informasi yang sudah tersimpan dalam ingatannya, selanjutnya menghubungkan atau mengubahnya serta menyampaikan informasi tersebut untuk mencapai tujuan atau menemukan kemungkinan jawaban dalam situasi yang membingungkan (Lewis dan Smith, 1993).

Penerapan yang sukses dari kemampuan berpikir tingkat tinggi terjadi ketika siswa berhasil menjelaskan, memutuskan, menunjukkan, dan menghasilkan penyelesaian masalah dalam konteks pengetahuan dan pengalaman (King, et.al, 2010). Keterampilan berpikir tingkat tinggi, harus dapat diukur dengan *assessment* yang jelas, valid, dan terkoordinasi sehingga hasilnya dapat dipercaya. Pengembangan instrumen evaluasi untuk mengukur keterampilan berpikir tingkat tinggi belum banyak dilakukan oleh praktisi pendidikan. Penilaian formatif yang ada sekarang ini hanya sedikit memberikan kesempatan pada siswa untuk mengembangkan pengetahuan lebih mendalam (Cullinane, 2011). Instrumen evaluasi yang mampu mengukur keterampilan berpikir tingkat tinggi mempunyai beberapa indikator antara lain: cenderung kompleks, memiliki solusi yang mungkin lebih dari satu (*open-ended approach*), dan membutuhkan usaha untuk menemukan struktur dalam ketidakteraturan (Lewi, 2009).

No	Indikator	Skor (%)	Kriteria
1	Konsep materi soal benar	100	Sangat baik
2	Cakupan materi sesuai tingkatan siswa	100	Sangat baik
3	Istilah yang digunakan jelas	97,5	Sangat baik
4	Materi soal mudah dipahami	92,5	Sangat baik
5	Materi soal ditulis sistematis, runtut, dan alur logika jelas	93,75	Sangat baik
	Rata-rata	**96, 75**	**Sangat baik**

Pengembangan instrumen evaluasi *Two-tier Multiple Choice Question* (*TT MCQ*) didasarkan pada teori perkembangan kognitif dari Piaget. Implikasi dari teori Piaget adalah instrumen yang dikembangkan disesuaikan dengan tingkat perkembangan kognitif siswa sehingga tidak terlalu sulit untuk dipahami. Bentuk instrumen yang dikembangkan sesuai dengan teori bepikir "John Dewey". Implikasi teori Dewey dalam pengembangan instrumen evaluasi adalah soal yang diberikan berupa masalah yang bertujuan untuk merangsang siswa meningkatkan kemampuan berpikir yang tidak hanya sekedar menghapal. Indikator instrumen evaluasi yang dikembangkan sesuai dengan teori kognitif Bloom yang telah direvisi oleh Anderson dan Kratwohl (2001) meliputi kemampuan menganalisis, mengevaluasi, dan menciptakan.

Pendapat John Dewey sejalan dengan teori konstruktivistik. Implikasi teori kontruktivistik dalam pengembangan instrumen evaluasi *two-teir multiple choice question* adalah instrumen evaluasi mengandung masalah yang harus dipecahkan siswa, untuk memecahkan masalah tersebut siswa harus memiliki keterampilan yang mengaitkan pengetahuan lama dengan pengetahuan baru. Keterampilan mengaitkan pengetahuan lama dengan pengetahuan baru tidak hanya membutuhkan keterampilan mengingat saja tetapi membutuhkan keterampilan lain seperti menganalisis, mengevaluasi, dan menciptakan. Hasil validasi ahli materi mengenai penilaian instrumen evaluasi *two-tier multiple choice question* (*TT-MCQ*) disajikan pada Tabel 5.

Tabel 5 Hasil Penilaian Indikator Materi *TT-MCQ*

Tabel 5 menunjukkan bahwa persentase rata-rata penilaian indikator materi yang ada pada soal adalah 96,75%

atau "sangat baik". Penilaian oleh ahli materi bertujuan untuk menjamin validitas isi dari instrumen evaluasi pengembangan. Perbaikan telah dilakukan sesuai saran dari ahli materi meliputi konsep materi soal, penyederhanaan penulisan soal, dan penulisan kunci jawaban. Hasil validasi ahli instrumen evaluasi mengenai penilaian instrumen evaluasi disajikan pada Tabel 6.

Tabel 6 Hasil Penilaian Indikator Konstruk Instrumen *TT-MCQ*

No	Indikator	Skor (%)	Kriteria
1	Butir soal sesuai indikator	98,75	Sangat baik
2	Butir soal sesuai dengan materi yang diajarkan	98,75	Sangat baik
3	Isi materi yang ditanyakan sesuai tingkatan siswa	98,75	Sangat baik
4	Soal hanya mengandung satu jawaban benar	97,50	Sangat baik
5	Pokok soal dirumuskan dengan jelas	85,00	Baik
6	Pokok soal merupakan kalimat yang diperlukan saja	86,30	Baik
7	Pilihan jawaban homogeny	83,80	Baik
8	Panjang alternatif pilihan jawaban sama	93,80	Sangat baik
9	Pokok soal tidak menunjuk ke arah jawaban yang benar	90,00	Sangat baik
10	Tidak ada kalimat "semua jawaban benar" atau "semua jawaban salah"	100	Sangat baik
11	Ditraktor atau pengecoh berfungsi	86,30	Baik
12	Letak pilihan jawaban benar ditentukan secara acak	92,50	Sangat baik
13	Pokok soal tidak mengandung pernyataan negatif ganda	91,00	Sangat baik
14	Wacana, gambar, atau grafik berfungsi	93,00	Sangat baik
15	Antara butir soal tidak tergantung satu sama lain	93,80	Sangat baik
16	Rumusan kalimat komunikatif	85,00	Baik
17	Kalimat menggunakan bahasa yang baik dan benar	86,00	Baik
18	Rumusan kalimat tidak mengandung penafsiran ganda	95,00	Sangat baik
19	Menggunakan bahasa yang umum (bukan bahasa lokal)	94,00	Sangat baik
20	Rumusan soal tidak mengandung pernyataan yang menyinggung perasaan	100	Sangat baik
21	Rubrik penilaian benar	92,50	Sangat baik
22	Penskoran objektif	95,00	Sangat baik
	Rata-rata	**92,56**	**Sangat baik**

Tabel 6 menunjukkan bahwa persentase rata-rata penilaian indikator konstruk yang ada pada soal adalah 92,56% atau "sangat baik". Penilaian oleh

ahli instrumen evaluasi bertujuan untuk menjamin validitas konstruk instrumen evaluasi pengembangan. Perbaikan telah dilakukan sesuai saran dari ahli instrumen evaluasi meliputi koherensi antara stem soal dengan alasan jawaban, kesesuaian taksonomi Bloom dalam soal, penyederhanaan stem soal, dan perbaikan kesalahan tata tulis. Hasil validasi instrumen evaluasi pada guru senior disajikan pada Tabel 7.

Tabel 7 Hasil Penilaian Indikator Kelayakan Instrumen *TT-MCQ*

No	Indikator	Skor (%)	Kriteria
1	Soal sesuai dengan KD	98,75	Sangat baik
2	Soal sesuai dengan indikator	97,50	Sangat baik
3	Soal dapat mengukur kemampuan berpikir tingkat tinggi	98,75	Sangat baik
4	Maksud pertanyaan jelas	73,13	Baik
5	Perintah mengerjakan soal jelas	73,13	Baik
6	Istilah yang digunakan jelas	96,86	Sangat baik
7	Susunan kalimat baik	73,13	Baik
8	Tidak ada kesalahan tata tulis, ejaan, dan tanda baca	96,88	Sangat baik
9	Kunci jawaban benar	100	Sangat baik
10	Penskoran objektif	87,50	Sangat baik
11	Waktu siswa cukup untuk mengerjakan soal	87,50	Sangat baik
	Rata-rata	**89,34**	**Sangat baik**

Tabel 7 menunjukkan bahwa persentase rata-rata penilaian indikator kelayakan instrumen evaluasi adalah 89,34% atau "sangat baik". Penilaian oleh guru senior bertujuan untuk mengetahui kelayakan instrumen evaluasi sebelum diterapkan di sekolah. Perbaikan telah dilakukan sesuai saran dari guru senior meliputi alokasi waktu pengerjaan soal, penulisan alasan jawaban soal, serta perbaikan kesalahan tata tulis.

Mardapi (2008) mengemukakan waktu yang dibutuhkan untuk mengerjakan tes bentuk pilihan ganda adalah 2-3 menit untuk setiap butir tes. Sukardjo (2008) dalam Salirawati (2011) menyatakan ujian selama 90 menit jumlah butir tes pilihan ganda sekitar 20-30 soal, yang berarti setiap butir soal dikerjakan selama 3-3,6 menit. Arikunto (2007) menyatakan bahwa alokasi waktu pengerjaan sebuah tes tergantung pada banyaknya butir tes dan bentuk soalnya. Berdasarkan kajian yang dilakukan maka waktu yang diberikan untuk mengerjakan soal *two-tier multiple choice question* adalah 60 menit untuk 20 soal yang diberikan, sehingga masing-masing butir soal dikerjakan selama 3 menit. Uji terbatas dilakukan melalui dua tahap yakni uji coba satu-satu dan uji coba kelompok kecil. Hasil penilaian uji coba terbatas disajikan pada Tabel 8.

Tabel 8 Hasil Penilaian Instrumen *TT MCQ* pada Uji Coba terbatas

No	Indikator	Skor (%)	Kriteria
1	Susunan kalimat	73,21	Baik
2	Maksud pertanyaan	75,00	Baik
3	Istilah yang digunakan	76,79	Baik
4	Perintah mengerjakan soal	76,79	Baik
5	Tidak ada kesalahan tata tulis, ejaan, dan tanda baca	76,79	Baik
	Rata-rata	**75,71**	**Baik**

Tabel 8 menunjukkan persentase rata-rata dari penilaian instrumen evaluasi oleh siswa pada uji coba terbatas adalah 75, 71% atau dinilai "baik". Uji coba terbatas bertujuan untuk mengetahui keterbacaan instrumen evaluasi yang dikembangkan. Perbaikan telah dilakukan sesuai saran dari siswa meliputi perbaikan skema dan gambar yang tidak jelas serta

perbaikan kesalahan tata tulis yang masih ada di beberapa soal.

Tanggapan siswa terhadap produk pengembangan antara lain soal pilihan ganda bertingkat lebih menantang dari pada soal pilihan ganda biasa, soal pilihan ganda bertingkat lebih mampu mengukur kemampuan berpikir tingkat tinggi siswa serta lebih menguji pemahaman siswa pada materi yang diberikan, soal pilihan ganda bertingkat mampu mengurangi siswa untuk menebak jawaban seperti pada pilihan ganda biasa.

Uji coba terbatas dilanjutkan dengan uji coba lapangan. Uji coba lapangan dilakukan untuk mengetahui validitas, reliabilitas, tingkat kesukaran, dan daya beda dari masing-masing butir soal yang dikembangkan. Pengujian butir soal ini dilakukan pada 64 orang siswa dari SMA Negeri 1 Gemolong. Rangkuman hasil pengujian butir soal pengembangan disajikan pada Tabel 9.

Tabel 9 Rangkuman Hasil Pengujian Instrumen *TT-MCQ*

No soal	Keputusan	Interpretasi	Tingkat kesukaran	Daya pembeda
1	Valid setelah revisi	Cukup	Mudah	Cukup
2	Valid	Cukup	Sedang	Tinggi
3	Valid setelah revisi	Cukup	Sedang	Cukup
4	Valid	Cukup	Sedang	Tinggi
5	Valid	Cukup	Sulit	Tinggi
6	Valid setelah revisi	Cukup	Sedang	Tinggi
7	Valid setelah revisi	Cukup	Mudah	Tinggi
8	Valid	Tinggi	Sedang	Sangat tinggi
9	Valid	Cukup	Sedang	Tinggi
10	Valid setelah revisi	Cukup	Sedang	Cukup
11	Valid	Cukup	Sedang	Cukup
12	Valid	Cukup	Sedang	Tinggi
13	Valid	Cukup	Sedang	Tinggi
14	Valid	Tinggi	Sedang	Sangat tinggi
15	Valid	Cukup	Sedang	Tinggi
16	Valid	Cukup	Sedang	Tinggi
17	Valid	Cukup	Mudah	Cukup
18	Valid	Cukup	Sedang	Cukup
19	Valid	Tinggi	Sedang	Sangat tinggi
20	Valid	Cukup	Sedang	Tinggi

Tabel 9 menunjukkan instrumen evaluasi hasil pengembangan memiliki memiliki karakteristik antara lain memilki validitas dengan interpretasi berkisar "cukup" sampai dengan "tinggi", memiliki reabilitas yang tinggi, memiliki tingkat kesukaran soal dengan proporsi 15% mudah: 80% sedang: 5% sulit, memiliki daya pembeda soal dengan interpretasi berkisar "cukup" sampai dengan "sangat tinggi".

Purwanto (2010) mengungkapkan bahwa sebuah tes yang dapat dikatakan baik sebagai alat pengukuran jika memenuhi persyaratan kualitas tes, yaitu memiliki validitas, reliabilitas, objektivitas, dan praktibilitas yang baik. Kepraktisan instrumen evaluasi didapatkan dari hasil penilaian oleh guru pengguna di sekolah. Hasil penilaian

kepraktisan instrumen evaluasi disajikan pada Tabel 10.

Tabel 10 Analisis Indikator Penilaian Kepraktisan Instrumen *TT-MCQ*

No	Indikator	Skor (%)	Kriteria
1	Biaya penyusunan tes terjangkau	62,50	Cukup
2	Waktu penyusunan tes tidak lebih dari 1 bulan	62,50	Cukup
3	Penyusunan tes dapat dilakukan guru biologi	87,50	Sangat baik
4	Penilaian tes mudah	87,50	Sangat baik
5	Mengolah hasil tes mudah	87,50	Baik
6	Pelaksanaan tes mudah	75,00	Baik
7	Waktu siswaa untuk pelaksanaan tes di sekolah cukup	75,00	Baik
	Rata-rata	**76,79**	**Baik**

Tabel 10 menunjukkan bahwa persentase rata-rata penilaian indikator kepraktisan instrumen evaluasi adalah 76,79% atau dinilai baik. Kepraktisan instrumen evaluasi adalah kemungkinan suatu instrumen evaluasi digunakan kembali oleh guru untuk mengukur tujuan pembelajaran pada suatu saat nanti (Purwanto, 2010). Produk pengembangan berupa instrumen evaluasi *two-tier multiple choice question* yang valid dan reliable selanjutnya dikorelasikan dengan instrumen evaluasi *multiple choice question* untuk mengukur keterampilan berpikir tingkat tinggi. Pengujian korelasi bertujuan untuk mendapatkan data tentang respon siswa terhadap penggunaan instrumen evaluasi bentuk *two-tier multiple choice question* di sekolah. Pengujian dilakukan pada dua kelas di SMA Negeri 3 Surakarta yang setara. Kedua kelas tersebut diambil dengan menggunakan teknik *cluster random sampling*. Kedua kelas diberi perlakuan yang sama, namun pada akhir pembelajaran satu kelas (X.2) diuji dengan tes bentuk *multiple choice question* dan kelas lain (X.5) diuji dengan tes bentuk *two-tier multiple choice question*. Hasil kedua perlakuan selanjutnya dihitung nilai korelasinya menggunakan rumus pearson. Data disajikan pada Tabel 11.

Tabel 11. Nilai Korelasi Keterampilan Berpikir Tingkat Tinggi pada Instrumen Evaluasi Bentuk *MCQ* dan Instrumen Bentuk *TT-MCQ*

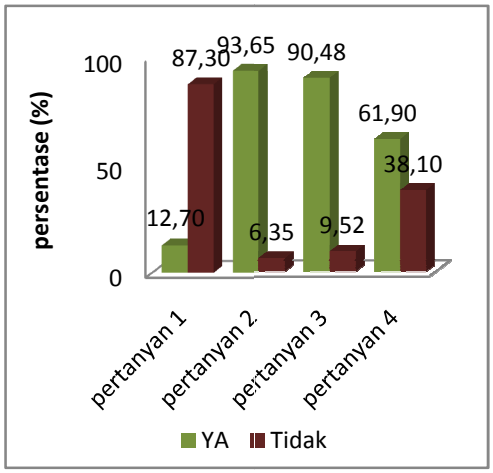
		Kelas X5	Kelas X2
Kelas X.5	Korelasi pearson	1	0,425
	Signifikansi		0,015
	N	32	32
Kelas X2	Korelasi pearson	0,425	1
	Siginifikansi	0,015	
	N	32	32

Korelasi signifikan kurang dari 0,05*

Hasil uji korelasi menunjukkan terdapat hubungan atau korelasi penerapan instrumen evaluasi antara bentuk *two-tier multiple choice question* (TTMCQ) dengan *multiple choice question* (MCQ), artinya siswa memberikan respon yang sama dalam mengerjakan soal, baik menggunakan instrumen evaluasi TTMCQ maupun menggunakan bentuk MCQ. Nilai korelasi menunjukkan hubungan yang positif, artinya jika siswa mampu mengerjakan tes dengan bentuk *multiple choice question* (MCQ) maka siswa juga mampu mengerjakan tes dengan bentuk *two-tier multiple choice question*.

Besarnya hubungan korelasi ditunjukkan dengan nilai sebesar 0,425, artinya respon siswa terhadap kedua bentuk instrumen evaluasi memiliki nilai yang cukup. Kesimpulan yang dapat diambil dari pengujian korelasi tersebut adalah siswa memberikan respon yang sama terhadap kedua bentuk instrumen evaluasi. Pengujian korelasi penting sebab instrumen TTMCQ merupakan instrumen yang baru diterapkan di sekolah tersebut. Gambar respon siswa terhadap penggunaan instrumen evaluasi *two-tier*

multiple choice question disajikan pada Gambar 1.



Gambar 1. Respon siswa terhadap penggunaan instrumen evaluasi *two-tier multiple choice question*

Keterangan :

- Pertanyaan 1: apakah kalian sebelumnya pernah diberikan tes dalam bentuk soal pilihan ganda bertingkat?
- Pertanyaan2: apakah bentuk soal pilihan ganda bertingkat lebih menantang daripada bentuk soal pilihan ganda biasa?
- Pertanyaan 3: apakah bentuk soal pilihan ganda bertingkat lebih dapat mengukur dan meningkatkan kemampuan berpikir daripada pilihan ganda biasa?
- Pertanyaan 4: apakah dikemudian hari kalian bersedia menggunakan soal tes pilihan ganda bertingkat untuk mengukur kemampuan berpikir?

Bentuk instrument	N	Rata-rata	Standar Deviasi	Standar eror Rata-rata
MCQ	32	70.09	11.061	1.955
TT-MCQ	32	65.69	16.155	2.856

Analisis terhadap Gambar 1 menunjukkan 87,30% siswa belum pernah menggunakan soal pilihan ganda bertingkat, artinya sebagian besar siswa masih belum mengenal bentuk soal pilihan ganda bertingkat (*two-tier multiple choice question*); 93, 65% siswa mengatakan bentuk soal pilihan ganda bertingkat lebih menantang daripada bentuk soal pilihan ganda biasa; 90, 48% siswa mengatakan bentuk soal pilihan ganda bertingkat lebih mengukur dan meningkatkan kemampuan berpikir dibandingkan dengan soal pilihan ganda biasa; 61, 90% siswa bersedia menggunakan soal tes pilihan ganda bertingkat untuk mengukur kemampuan berpikir.

Respon siswa terhadap instrumen evaluasi hasil pengembangan sesuai dengan pendapat dari Halaydina dan Downing (1989) serta Treagust (2006). Penelitian Haladyna dan Downing (1989) menyebutkan bentuk soal bentuk *two-tier multiple choice question* dapat digunakan untuk menguji pemahaman siswa serta mengukur keterampilan kognitif pada level yang lebih tinggi (*higher order thinking*). Treagust (2006) menyebutkan soal bentuk *two-tier multiple choice question* dapat digunakan untuk meningkatkan kemampuan berpikir siswa. Penelitian lain yang mendukung adalah penelitian dari Cullinane (2011) yang menyebutkan bahwa penggunaan bentuk *two-tier multiple choice question* mampu meningkatkan penilaian pembelajaran dan keterampilan berpikir yang lebih mendalam.

Rata-rata nilai keterampilan berpikir pada siswa yang diuji dengan TTMCQ dan MCQ digunakan untuk mengetahui tingkat keberhasilan siswa dalam mencapai tujuan pembelajaran ditunjukkan pada Tabel 12.

Tabel 12. Rata-rata Nilai Instrumen Evaluasi Bentuk *MCQ* dan Instrumen Bentuk *TT-MCQ*

Nilai rata-rata keterampilan berpikir tingkat tinggi siswa pada kedua bentuk instrumen menunjukkan nilai dibawah

KKM, artinya keterampilan berpikir tingkat tinggi siswa belum menunjukkan hasil yang maksimal. Kemampuan berpikir tidak dapat terjadi secara spontan karena kemampuan ini perlu untuk dilatihkan. Pengubahan kemampuan berpikir seseorang dibutuhkan sebuah proses dan latihan yang tidak singkat (Afcariono, 2008; Richomond, 2007; Wolf *et.al.*, 2005). Belajar untuk mengembangkan keterampilan berpikir akan berhasil apabila banyak dilakukan latihan atau ulangan (Sagala, 2011).

Faktor yang mempengaruhi hasil belajar siswa termasuk keterampilan berpikir tingkat tinggi dapat berasal dari faktor eksternal seperti lingkungan keluarga dan lingkungan sekolah maupun faktor internal seperti kondisi fisiologis dan psikologis siswa (Suryabrata, 2005). *Law of readiness* menyatakan bahwa apabila satuan-satuan dalam system syaraf telah siap berkonduksi dan hubungan itu berlangsung atau dengan kata lain siswa telah siap menerima *stimulus* atau rangsangan pelajaran, maka terjaminnya hubungan antara *stimulus* dengan tanggapan siswa akan memuaskan. Hubungan stimulus-respon akan terbentuk dan melahirkan tingkah laku baru apabila siswa telah siap belajar (Sagala, 2011).

Temuan di lapangan menunjukkan bahwa soal hasil pengembangan dalam bentuk *two-tier multiple choice question* memiliki keunggulan dan kelemahan. Keunggulan soal pilihan ganda bertingkat antara lain jumlah materi yang dapat ditanyakan relatif lebih banyak dibandingkan dengan materi yang dicakup soal bentuk uraian; dapat mengukur jenjang kemampuan berpikir tingkat tinggi (analisis, evaluasi, mencipta) yang umumnya sulit dilakukan oleh soal pilihan ganda biasa; penskoran mudah, cepat, dan objektif; reliabilitas soal relatif lebih tinggi dibandingkan dengan soal uraian; dapat digunakan untuk mengukur kemampuan *problem solving*; dapat digunakan sebagai alat diagnosis pemahaman materi siswa; dapat digunakan untuk mendeteksi miskonsepsi yang mungkin dimiliki siswa; dapat digunakan untuk mengetahui efektifitas pembelajaran yang dilakukan guru; peluang untuk menerka atau menembak jawaban lebih sedikit karena antara soal tingkat pertama dengan soal tingkat kedua saling berkait. Kelemahan soal pilihan ganda bertingkat antara lain kurang dapat digunakan untuk mengukur kemampuan verbal; penyusunan soal yang baik memerlukan waktu yang relatif lama dibandingkan dengan bentuk soal yang lainnya; siswa belum terbiasa menggunakan soal dalam bentuk pilihan ganda bertingkat (*TT-MCQ)*; guru belum pernah menggunakan soal pilihan ganda bertingkat (*TT-MCQ).*

Kesimpulan dan Saran

Kesimpulan dari penelitian pengembangan evaluasi antara lain:

1. Karakteristik instrumen evaluasi *two-tier multiple choice question* yang mengukur keterampilan berpikir tingkat tinggi antara lain dikembangkan berdasarkan indikator keterampilan berpikir tingkat tinggi dari Anderson dan Krathwohl (2001) meliputi keterampilan menganalisis, mengevaluasi, serta menciptakan, memiliki validitas dengan interpretasi minimal "cukup", serta memiliki reabilitas yang tinggi.
2. Kelayakan produk instrumen evaluasi *two-tier multiple choice question* dijamin melalui validitas isi yang dinilai baik oleh ahli materi, validitas konstruk yang dinilai baik oleh ahli instrumen evaluasi, validitas butir soal dengan interpretasi minimal cukup, memiliki tingkat kesukaran soal dengan proporsi 15% mudah: 80% sedang: 5% sulit, memiliki daya pembeda soal dengan interpretasi minimal "cukup", serta memiliki

tingkat kepraktisan soal yang dinilai baik.

3. Respon siswa terhadap penerapan instrumen evaluasi *two-tier multiple choice question* didapatkan melalui hasil angket respon siswa terhadap penerapan instrumen evaluasi serta uji korelasi antara instrumen bentuk *two-tier multiple choice question* dengan bnetuk *multiple choice question*. Hasil uji korelasi menunjukkan ada korelasi antara kedua bentuk instrumen tersebut dengan nilai sebesar 0,15, artinya siswa memberikan respon yang sama dalam mengerjakan soal baik menggunakan instrumen evaluasi *two-tier multiple choice question* maupun menggunakan bentuk *multiple choice question*.

Rekomendasi untuk penelitian pengembangan evaluasi antara lain:

1. Siswa sebelumnya telah mendapatkan materi kingdom plantae yang termasuk KD mendeskripsikan ciri-ciri divisio dalam dunia tumbuhan dan peranannya bagi kelangsungan hidup di bumi.
2. Guru sebaiknya membelajarkan materi pada KD tersebut menggunakan model pembelajaran yang memberdayakan keterampilan berpikir tingkat tinggi.
3. Alokasi waktu yang diberikan untuk mengerjakan 20 soal bentuk *two-tier multiple choice question* tidak kurang dari 60 menit.
4. Evaluasi yang terkait dengan model pembelajaran yang memberdayakan keterampilan berpikir tingkat tinggi masih perlu dikembangkan dan diteliti lebih lanjut.

Daftar Pustaka

Anderson, L.W dan D.R Krathwohl. (2001). *A Taxonomy for Learning, Teaching, and Assesing*. New York: Longman

Ball, Anna L dan Bryan L. Garton. (2005). Modelling Higher Order Thinking: The Allignment Between Objectives, Classroom Discourse, and Assesment. *Journal of Agricultural Education*, Volume 46, Number 2, 2005.

Borg, W.R & Gall, M.D. (1983). *Educational Research An Introduction (4th Ed)*. White Plains: Logman Inc.

Cullinane, Alison dan Maeve Liston. (2011). *Two-tier Multiple Choice Question: An Alternative Method of Formatif Assessment for First Year Undergraduate Biology Students*. Limerick: National Center for Excellence In Mathematics and Education Science Teaching and Learning (NCE-MSTL).

Das Salirawati,. (2011). Pengembangan Instrumen Pendeteksi Miskonsepsi Kesetimbangan Kimia pada Peserta Didik SMA. *Jurnal Pendidikan dan Evaluasi Pendidikan* Tahun 15, Nomor 2, 2011.

Djaali dan Pudji Mulyono. (2008). *Pengukuran dalam Bidang Pendidikan*. Jakarta: Grasindo.

Djemari Mardapi. (2008). *Teknik Penyusunan Instrumen Tes dan Nontes*. Yogyakarta: Mitra Cendekia Yogyakarta Press.

Halaydina, T.M dan Downing, S.M. (1989). *A Taxonomy of Multiple Choice Item Writing Rules*. Applied Measurements In Education, 2(1), 37-50.

King, JF; Goodson, Ludwika, dan Rohani, Faranak. (2010). *Higher Order Thinking Skills, Definition, Teaching Strategis, Assesment.* A Publication of The Educational Services Program. Tersedia di www.Cala.fsu.edu

Lewis, A & Smith, D. (1993). *Defining Higher Order Thinking*. Theory Into Practice, 32(3), 131-137

Muchamad Afcariono. (2008). Penerapan Pembelajaran Berbasis Masalah untuk Meningkatkan Kemampuan Berpikir Siswa pada Mata Pelajaran Biologi. *Jurnal Pendidikan Inovatif*. 3(2): 65-68

Ngalim Purwanto. (2010). *Prinsip-prinsip dan Teknik Evaluasi Pengajaran.* Bandung: PT Remaja Rosdakarya.

Permendiknas No 23. (2006). *Standar Kompetensi Lulusan untuk Satuan Pendidikan Dasar dan Menengah*. Jakarta: Depdiknas.

Permendiknas No 20. (2007). *Standar Penilaian Pendidikan*. Jakarta: Depdiknas.

Pohl. (2002). *Learning Thinking to learn.* tersedia di www.purdue.edu/geri

Ramirez, Rachel Patricia B dan Mildred S. Ganaden. (2006). Creative Activities and Students' Higher Order Thinking Skills. *Journal of Education Quarterly*, December 2008, vol 66 (1), 22-23.

Richmond, Jonathan E.D. (2007). Bringing Critical Thinking to The Education of Developing Country Professionals. *International Education Journal,* 2007, 8(1), 1-29.

Syaiful Sagala,. (2009). *Konsep dan Makna Pembelajaran: untuk Membantu Memecahkan Masalah Belajar dan Mengajar*. Bandung: Alfabeta.

Sajidan. (2012). Penerapan Model Pengembangan Mutu Pendidikan dalam Rangka Peningkatan Kompetensi Guru SMA Melalui Pengembangan *Subject Specific Paedagogy* (SSP). *Draft Artikel Penelitian*. Universitas Sebelas Maret Surakarta.

Suharsimi Arikunto. (2011). *Dasar-dasar Evaluasi Pendidikan*. Jakarta : PT. Bumi Aksara

Sumadi Suryabrata,. (2005). *Psikologi pendidikan*. Bandung: Rajawali Pers.

Treagust, David F. (2006). Diagnostic Assesment In Science as A Means to Improving Teaching, Learning, and Retention. *UniServe Science Assesment Symposium Proceedings*. The University of Sydney, 28 September 2006.

Weiss, Renée E. (2003). Designing Problems to Promote Higher Order Thinking. *New Direction for Teaching and Learning*, No 95, Fall 2003.